RESEARCH-ARTICLE

Higher-Order Vision-Language Alignment for Social Media Prediction

MINGSHENG TU, Chongqing University of Posts and Telecommunications, Chongqing, Chongqing, China

TIANJIAO WAN, National University of Defense Technology China, Changsha, Hunan, China

QISHENG XU, National University of Defense Technology China, Changsha, Hunan, China

XINHAO JIANG, Fujian Agriculture and Forestry University, Fuzhou, Fujian, China

KELE XU, National University of Defense Technology China, Changsha, Hunan, China

CHENG YANG, National University of Defense Technology China, Changsha, Hunan, China

Published: 28 October 2024

MM '24: The 32nd ACM International Conference on Multimedia
October 28 - November 1, 2024
Melbourne VIC, Australia

Conference Sponsors:
SIGMM

https://doi.org/10.1145/3664647.3688999
ISBN: 9798400706868

Higher-Order Vision-Language Alignment for Social Media Prediction

Mingsheng Tu*
Chongqing University of Posts and Telecommunication
Chongqing, China

Tianjiao Wan*
National University of Defense Technology
Changsha, Hunan, China

Qisheng Xu*
National University of Defense Technology
Changsha, Hunan, China

Xinhao Jiang
Fujian Agriculture and Forestry University
Fuzhou, Fujian, China

Kele Xu†
National University of Defense Technology
Changsha, Hunan, China

Cheng Yang
National University of Defense Technology
Changsha, Hunan, China

Abstract

The prediction task of social media popularity aims to automatically forecast the future popularity of the posts by leveraging vast amounts of social media data. This data encompasses diverse visual and textual content, including photos, categories, custom tags, temporal information, and geographical data. Existing methods have explored multiple feature types to enhance popularity prediction. Despite their success, visual and textual features—both crucial pieces of information—are often simply concatenated after extraction, ignoring the divergence between these two feature spaces. In this paper, we propose a method to project visual and language information into an aligned semantic representation, thereby uncovering intricate associations between these two modalities. Specifically, we leverage the BLIP-2 model to understand and generate visual description text that encapsulates the content of photos. Semantic embeddings are then extracted from all available visual and textual information. Additionally, we deeply exploit user-related behavior and characteristic information to extract features, uncovering hidden clues for post popularity prediction. Leveraging these improvements, we conduct extensive experiments to demonstrate the effectiveness of our proposed method.

CCS Concepts

• **Human-centered computing** → **Collaborative and social computing theory, concepts and paradigms**; **Social networks**.

Keywords

Multimodal Machine Learning, Social Media Popularity Prediction

*These authors contributed equally.
†Corresponding author, kl_xu2024@outlook.com

MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia

ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3688999

ACM Reference Format:
Mingsheng Tu, Tianjiao Wan*, Qisheng Xu*, Xinhao Jiang, Kele Xu, and Cheng Yang. 2024. Higher-Order Vision-Language Alignment for Social Media Prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3664647.3688999

1 Introduction

With the rapid development of the Internet, massive amounts of content are generated daily on social media platforms like Twitter, Instagram, and Flickr. However, not all posts receive equal attention. Analyzing and predicting the popularity of these posts can enhance user experiences by delivering popular information to them [30]. Therefore, the Social Media Prediction (SMP) task [37] has been proposed to accurately predict the future popularity of each online post using social media data. This can be applied to various real-world applications, including online advertising [10], information retrieval [8] and social recommendations [24].

The SMP task poses a significant challenge due to the diverse nature of social media data, which encompasses various types of information, including visual content and textual content (e.g., posted photos, custom tags, spatial locations, user profiles, and other forms). Taking all of this multimodal information into account and achieving precise social media predictions can be difficult. Previous methods [41] have explored many meaningful and significant approaches to feature extraction, feature combination, and importance weighting. Despite sustained efforts in the field of SMP, it is crucial to acknowledge that most SMP research to date has primarily concentrated on simply extracting information from single-modal data, often ignoring the high-order interactions of multimodal features [1].

Specifically, the SMP task encounters the following challenges: Firstly, the difference between visual feature space and textual feature space is usually overlooked. The alignment of multimedia representations has always been a critical task in the field of multimedia processing, involving mapping features from different modalities into a common representation space to enhance overall understanding. In the context of the SMP task, the photos have a closely intertwined connection with the titles and tags in one post, which together play a significant role in predicting post popularity as shown in Figure 1. The common practice is to use deep neural networks to extract visual and textual features separately and then

Post(a) Photo	
Category	Whether&Season
Title	Deep blue sky
Alltags	Road, blue, trees, winter, light, sunset, red, sky, orange, sun, sunlight.....
Popularity	16.56

Post(a) Photo	
Category	Fashion
Title	One world
Alltags	Nyc, newyorkcity, blackandwhite, bw, white, black, byn, blanco
Popularity	16.45

Figure 1: Samples of social media posts. The visual content shows a strong correlation with the titles and tags in one post, the alignment of multimedia representations is essential for accurately predicting popularity.

concatenate them. It can be insufficient since these features often do not reside in a unified feature space. To enhance performance, achieving effective vision-language alignment is necessary [26]. Additionally, due to the complexity of the SMP task, user portrait information should also be emphasized to capture the hidden clues behind post popularity. Although previous research has been effective in extracting features from post profiles, it often overlooks the significant influence of user-related information on determining the popularity of social media posts. In fact, the popularity of a post depends not only on its content but also on the behaviors and characteristics of the users who publish it. For example, a post from a popular user who frequently uses trending tags is more likely to achieve higher engagement compared to posts from users with low-engagement tags.

To overcome the above two challenges encountered in the popularity prediction task, in this paper, we propose a novel Higher-order Vision-Language Alignment method for social media prediction, named "HVLA". To bridge the multimodal feature gap, the Bootstrapping Language-Image Pre-training model with frozen unimodal models (BLIP-2) is employed to generate photo descriptions guided by text, including titles and tags. By achieving a semantic understanding of the photo, this approach seamlessly integrates visual and textual information into an aligned semantic feature space. It avoids the information loss typically resulting from the separation of visual and textual features in traditional methods, thereby enhancing the effectiveness of multimodal data integration. In terms of tailored feature engineering, we utilize user behavior features and activity levels to provide more accurate insights for predicting popularity, alongside basic statistical features. Specifically, we track the sequences of tags from users' posts to identify dynamic behavior patterns, thus capturing the interest trends in the content they publish. Additionally, we calculate the total number of posts and photos, posting intervals, and post volume fluctuations to capture the user's activity level on the social platform.

The main contributions of this work are as follows:

- We propose a Higher-order Visual-language Alignment method for social media prediction, called HVLA. We introduce the BLIP-2 model to generate text-guided image descriptions, mapping multimodal information into common semantic feature space.
- Considering the user's impact on post popularity, we construct user-related features mainly based on dynamic behavior patterns and activity levels to more accurately predict popularity.
- Extensive experimental results prove the effectiveness of our framework in significantly improving the accuracy of predicting post popularity, while achieving second place in the Challenge.

2 Related Work

2.1 Social Media Popularity Prediction

Social media popularity prediction has become a crucial area of research due to the exponential growth of user-generated content and dynamic attributes [35]. Researchers have developed various models and techniques to predict the popularity of social media posts, focusing on different aspects such as feature extraction, user behavior, and network dynamics [34].

The authors in [34] propose to integrate multi-modal feature extraction for feature generalization and temporal modeling. They utilize two separate CatBoost models trained on distinct datasets to address data sparsity. Sliding window averages are also implemented to capture short-term dependencies in user post sequences. In [39], various types of features, including visual and language information, are deeply exploited to improve the reliability in modeling post popularity. They analyze images from multiple perspectives, while extracting semantic embeddings from diverse textual sources like titles, tags, concepts, and categories. TTC-VLT [3] improves prediction accuracy by utilizing contrastive learning between titles and tags while also embedding user identification features. Furthermore, [16] introduces a diverse feature mining method and a stacking block module to capture feature information from text and images, significantly improving prediction performance.

2.2 Multimodal Representation Learning

The effectiveness of popularity prediction relies on the advanced capabilities of multimodal representation [25], which accurately represents complex data across multiple types of modalities. The vision-language representation learning consists of a wide range of tasks, including visual question and answer, image captioning, image-text matching and visual reasoning, etc. Significant progress in natural language processing (NLP) [5, 31], image recognition [28], and other relevant technologies have spurred the continued advancement of this field.

The authors in [12] propose a framework to learn unified representations of images and text using adversarial training, fostering cross-modal translation and alignment. Moreover, BERT is extended in [13] to multimodal contexts, illustrating how pretraining can enhance performance across diverse vision-and-language tasks. Notably, models such as CLIP (Contrastive Language–Image Pre-training) [25] achieve results far exceeding previous works by scaling up visual-language pretrained models. Building on the success of CLIP, BLIP-2 [11] leverages off-the-shelf frozen pre-trained image encoders and large language models to bootstrap vision-language pretraining.

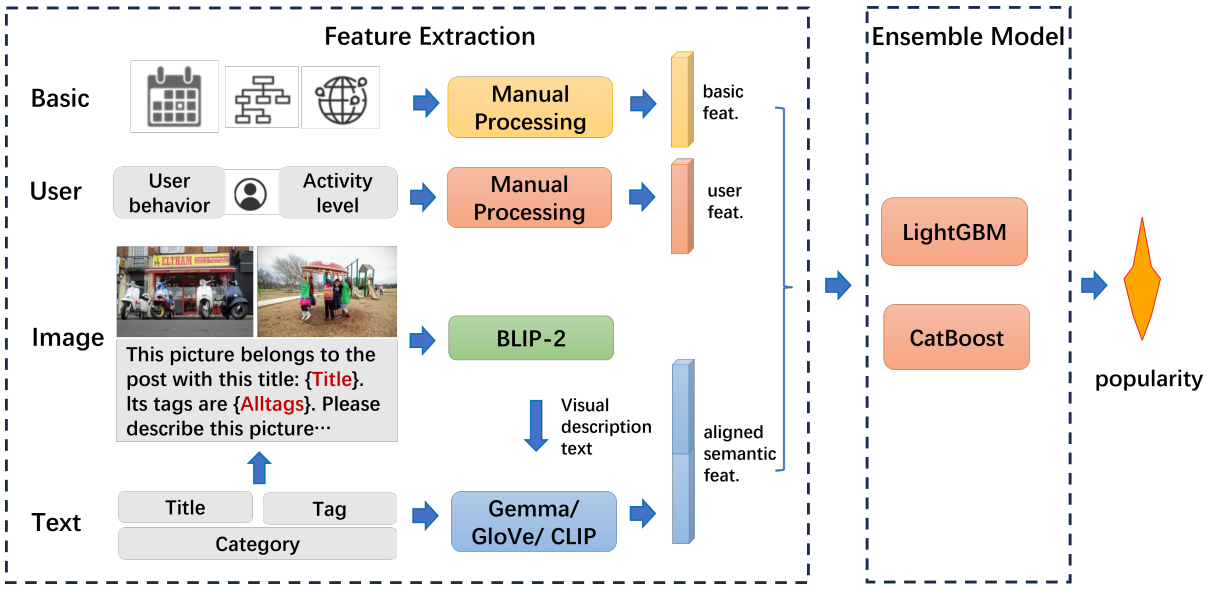

Figure 2: The overall framework of the proposed method involves several key steps. For the hand-crafted features, we extract hand-crafted features independently from the post's and user's perspective respectively to reveal latent information. For the high-order visual-language feature, we project them into an aligned semantic feature space using the Blip-model. Subsequently, all these diverse features are concatenated and fed into an ensemble prediction module, consisting of LightGBM and CatBoost models to output an accurate predicted popularity score.

2.3 Multimodal Feature Fusion

Multimodal feature fusion has garnered significant attention in recent years, particularly in the context of social media analysis, image captioning, and video understanding. Early approaches to multimodal fusion primarily relies on simple concatenation methods, where features from different modalities are extracted separately and then combined into a single feature vector [17, 33]. While these methods demonstrate the utility of combining visual and textual features for tasks like event detection and popularity prediction, they often fail to capture the complex interactions and dependencies between different modalities, leading to sub-optimal performance.

To address the limitations of simple concatenation, more advanced fusion techniques have been developed. Attention mechanisms, as introduced by [32], have been successfully applied to multimodal fusion tasks [41], allowing models to dynamically weight the importance of different modalities. For instance, [15] employ attention mechanisms to fuse visual and textual information for image captioning, significantly improving the quality of generated captions. Moreover, models like VILBERT [14] and LXMERT [29] introduce pre-trained multimodal transformers that jointly model visual and textual information, achieving state-of-the-art results in various vision-and-language tasks. These models utilize a two-stream architecture where visual and textual inputs are processed separately before being fused at multiple layers, capturing fine-grained interactions between the modalities.

3 Method

The overall framework is shown in Figure 2. We integrate the textual and image information in a post into an aligned semantic feature space. For the hand-crafted features, we extract rich features from the post's and user's perspective for improving the performance of popularity prediction (Section 3.2). Finally, in the Section 3.3, all these features are concatenated and fed into the ensemble prediction module to derive the predicted popularity score.

3.1 Higher-order Vision-Language Feature

In order to bridge the modality gap, we have employed the BLIP-2 model to generate visual description text firstly, considering both photos and text. Then the descriptions are processed with other textual information to extract unified semantic features.

3.1.1 Visual information extraction. Visual information can immediately capture users' attention and facilitate their understanding of post content. With textual information, they play a critical role in determining the post's popularity. However, there exists a significant disparity between visual feature space and textual feature space, leading to information loss when directly concatenating them.

To tackle this challenge, we utilize the BLIP-2 model to project both modalities into an aligned semantic vector space. As mentioned earlier, the BLIP-2 model [11] has achieved strong generalization ability across tasks such as image captioning, so we employ it to extract the descriptive details from the visual content. In detail, our approach begins by crafting a tailored prompt for each post based on its title and associated tags. Subsequently, the photo is

then inputted, and BLIP-2 [11] model is utilized to understand the content of the image guided by the prompt, thereby generating a comprehensive visual description text.

By harnessing the BLIP-2 model's capacity to comprehend and generate detailed descriptions that encapsulate the photo's content, the photo information is transformed into textual format, which allows for natural alignment with other textual elements of the post. This alignment enables the discovery of intricate associations between the textual and visual components, preventing disjointed interpretations. Such high-order multimodal information is pivotal for deeper insights into the visual-language nature of posts.

3.1.2 Textual information extraction. For the textual information extraction, the photo's descriptive text is processed in the same manner as the other textual information in the post. It includes titles, tags, etc., that generally describe the content the user wants to convey. Besides, other features derived from statistical analysis are also incorporated through manual processing. To enhance the diversity of text representation, we use multiple language models for feature extraction. First, we employ the advanced language model, Gemma [20], to extract diverse text embedding representations. With its 2 billion parameters, Gemma excels at handling complex textual features and thoroughly mining the rich information contained within the text. The second uses GloVe [22] model, which uses the global statistical information of a corpus to generate vectors. Additionally, we also utilize CLIP [25] model to enrich feature vectors.

3.2 Hand-crafted Features Extraction

For SMP task, various types of features can be manually constructed. In addition to the directly-extracted feature, the hand-crafted feature we explore includes two main perspectives: the post's perspective and the user's perspective. From the post's perspective, we analyze patterns over weekly and monthly time intervals, including the hot tags, the frequency of hot topics appearing, and the distribution of post contents across different tags [19]. Furthermore and most importantly, we believe that the post popularity also highly relies on the poster's personality, including dynamic behavior patterns, activity level, personal preference and others [4]. To deeply analyze these hidden clues behind post popularity, we construct a range of hand-crafted features from the user's perspective. First of all, to analyze the dynamic user behavior patterns, we track the sequence of tags used in their past posts. By examining these tags, we can gain insights into their posting behavior and interest areas. We also calculate the total number of tags, which indicates the user's engagement with the content and their tagging habits [18]. Additionally, to evaluate user activity levels and contribution to the platform, we count the number of posts and photos they share within various time intervals. We also analyze the intervals between posts published on the website to capture the user's posting frequency and activity patterns. Considering that the activity level of users changes over time, we further compare the number of photos posted in the first year with those in the last year, thus discerning changes in users' personal preferences [21, 27].

3.3 Ensemble Prediction Module

To boost accuracy in predicting social media popularity, we have developed an ensemble prediction module leveraging both CatBoost [23] and LightGBM [9]. In our methodology, the above mentioned features are all concatenated and fed into both CatBoost and LightGBM models. The outputs from these models are subsequently fused using a weighted fusion strategy to forecast the popularity score of the post. By integrating their predictive capabilities, our ensemble prediction module not only provides a more nuanced understanding of the data but also significantly enhances the robustness of our forecasting models [6, 42].

The fusion strategy is presented as follows:

$$P_{score} = \lambda \times P_{Catboost} + (1 - \lambda) \times P_{LightGBM} \tag{1}$$

where P_{score} donates the final predicted popularity score of the post, $P_{CatBoost}$ and $P_{LightGBM}$ donate the predicted scores assigned by the Catboost and LightGBM models, respectively. The aggregation of these scores is guided by a hyperparameter λ, which serves as a weighting factor to balance the contributions from each model's prediction.

4 Experiments

4.1 Dataset

The Social Media Prediction Dataset (SMPD) [35, 36, 38] is a large-scale benchmark, collected from a large photo sharing website Flickr. The overview statistics reveal it includes over 486k posts from 69k users. Each post contains both visual and textual content such as photos, categories, tags, temporal and geographic information. These attributes provide a robust platform for exploring the popularity prediction from online social media data.

4.2 Evaluation Metrics

Following [38], we leverage two popular metrics for quantitative evaluation of model performance: Spearman's Rho (SRC) and Mean Absolute Error (MAE) [7, 34]. SRC measures the relevance between predicted and ground-truth popularity ranks, while MAE focuses on the average accuracy error.

Spearman's Rho (SRC) is defined as:

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}, \tag{2}$$

where d_i is the difference between the ranks of corresponding values, and n is the number of observations. SRC evaluates the rank correlation, providing insights into how well the predictions preserve the ordering of the ground-truth values.

Mean Absolute Error (MAE) is calculated as:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i|, \tag{3}$$

where y_i is the ground-truth value, $\hat{y}_i$ is the predicted value, and n is the number of observations. MAE measures the average magnitude of the errors in a set of predictions, without considering their direction, providing a straightforward interpretation of model accuracy. These metrics will be calculated for each approach to ensure a fair performance evaluation [2, 40].

Table 1: Experimental results with various settings (Red: Best; Blue: Second Best). ↑ indicates better performance with higher values, while ↓ indicates better performance with lower values.

Base feat.	Glove	CLIP	BLIP-2	Gemma	Users Related features	CatBoost		LightGBM		Fusion	
						SRC ↑	MAE ↓	SRC ↑	MAE ↓	SRC ↑	MAE ↓
√						0.7023	1.2783	0.6833	1.2971	0.7042	1.2615
√				√		0.7219	1.2318	0.7058	1.2520	0.7223	1.2214
√			√			0.7223	1.2322	0.7135	1.2464	0.7285	1.2148
√		√				0.7250	1.2249	0.7161	1.2353	0.7296	1.2087
√		√		√		0.7241	1.2283	0.7157	1.2354	0.7292	1.2093
√	√					0.7245	1.2208	0.7110	1.2414	0.7276	1.2278
√	√	√		√		0.7273	1.2162	0.7177	1.2284	0.7319	1.2001
√	√	√	√	√		0.7300	1.2152	0.7190	1.2262	0.7330	1.1979
√					√	0.7058	1.2555	0.6970	1.2724	0.7127	1.2410
√				√	√	0.7225	1.2287	0.7112	1.2448	0.7258	1.2154
√			√		√	0.7237	1.2264	0.7127	1.2448	0.7269	1.2136
√			√	√	√	0.7226	1.2354	0.7184	1.2343	0.7294	1.2123
√	√	√	√	√	√	0.7295	1.2172	0.7203	1.2230	**0.7341**	**1.1970**

4.3 Overall Results

To validate the effectiveness of our proposed Higher-order Vision-Language Alignment method in the SMP task, two base regression models are employed to predict the post popularity in our experiments, including LightGBM [9] and CatBoost [23]. The overall experimental results are illustrated in Table 1. Below we systematically analyze the effectiveness of our proposed modules in model performance enhancement.

First of all, only the basic features in [38] are fed into the regression prediction model to achieve a baseline score. These baseline scores serve as a reference point for assessing the efficacy of subsequent model modules. With the only official given features, it's found that using the CatBoost model tends to yield improved results compared to using the LightGBM.

In Section 3.1, we bridge the modality gap by first employing a BLIP-2 model to generate visual description text. This description text, along with other textual information, is then processed to extract semantic features. For visual information extraction, introducing the BLIP-2 model resulted in improvements of 0.020 and 0.030 in SRC when using the CatBoost and LightGBM models, respectively. For textual information extraction, leveraging the Gemma model to align the description text and other textual information into semantic features, resulting in improvements of 0.019 and 0.023 in SRC, as shown in Table 1. To enhance text representation diversity, we introduce additional language models, GloVe and CLIP, to enrich the text features. It reduces the MAE score by 0.016 and 0.024, respectively, compared to using Gemma alone. Overall, aligning visual and textual information into semantic space improves the SRC score by 0.028 and 0.036, and the MAE score by 0.063 and 0.071.

In accordance with Section 3.2, user-related features, such as the sequence of tags and the total number of tags, have been exploited to provide hidden clues for post popularity prediction. We find that extra user-related features could improve the MAE metric with 0.023 and 0.025 using CatBoost and LightGBM respectively, demonstrating its effectiveness. This is also consistent with our motivation that user characteristics have a close connection with the post popularity. Consequently, analyzing the user's behavior pattern and activity level contributes to boosting our prediction accuracy.

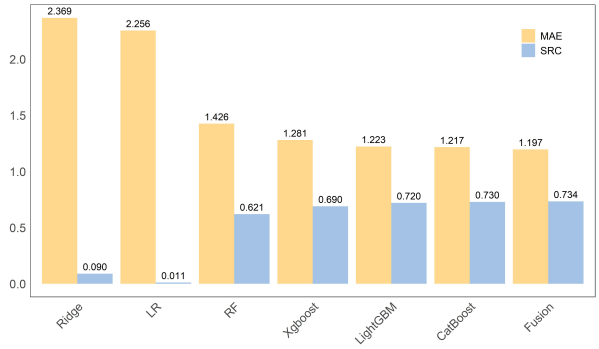

Figure 3: Compared performance of different models.

Finally, our method achieves an SRC score of 0.7295 and an MAE score of 1.2172 using the CatBoost model on the SMP dataset, while the LightGBM model achieves lower scores (0.7203 for SRC and 1.2230 for MAE). This indicates that utilizing hand-crafted user features and aligned multimodal features significantly improves our ability to predict the popularity of posts more accurately.

Despite the impressive accuracy achieved by deeply exploiting various types of features in our method, the model's performance can be further boosted by enhancing its robustness. As mentioned in Section 3.3, an ensemble method is utilized to consider the predictions from both CatBoost and LightGBM, thus providing a more comprehensive understanding of the posts. By fusing the predicted outputs, the SRC score increases from 0.7295 and 0.7203 (CatBoost and LightGBM) to 0.7341, and the MAE score decreases from 1.2172 and 1.2230 to 1.1970, highlighting the contribution of ensemble prediction module.

Rank	Features	Imp	Rank	Features	Imp
(1)	Uid_post_diff_min	364	(11)	Uid_alltags_numskew	150
(2)	Uid_pid_skew	203	(12)	Uid_title_length_var	143
(3)	Uid_Pid_min	199	(13)	Uid_pid_var	137
(4)	Uid_post_diff_std	191	(14)	Uid_alltags_lenskew	129
(5)	Uid_post_diff_mean	190	(15)	Uid_post_diff	125
(6)	Uid_post_diff_skew	181	(16)	Uid_alltags_lenstd	114
(7)	Uid_Pid_std	161	(17)	timezone_offset_Uidcount	108
(8)	Uid_title_length_std	160	(18)	Uid_alltags_lenva	100
(9)	Uid_postdate_skew	158	(19)	Uid_alltags_numvar	98
(10)	Uid_title_length_skew	152	(20)	photo_count	98

Table 2: Feature importance rankings

4.4 Performance Comparison

To evaluate prediction performance, several high-performing regression models are selected for comparison. The corresponding results are depicted in Fig. 3. Among the models, simple linear models like Ridge and LR exhibit relatively low SRC scores. In contrast, CatBoost and LightGBM demonstrate superior performance over other baselines, showcasing their effectiveness in social media prediction tasks. Consequently, our paper employs these two models for the SMP task. Furthermore, we enhance prediction robustness by employing an ensemble method that combines predictions from both CatBoost and LightGBM, which achieves the best performance.

4.5 Feature Importance

For the SMP task, we utilized LightGBM to identify the most influential features contributing to the prediction of post popularity (Table 2). The importance of each feature is measured by its contribution to reducing the loss (or impurity) in the decision trees. The feature importance in LightGBM is computed based on:

$$\text{Importance}(f_i) = \sum_{t \in T} \sum_{s \in S_t} \Delta L(s,t) \cdot \mathbb{I}(s \text{ uses } f_i), \tag{4}$$

where T is the set of all trees in the model, S_t is the set of all splits in tree t, $\Delta L(s,t)$ is the loss reduction at split s in tree t, and $\mathbb{I}(\cdot)$ is an indicator function that equals 1 if the split uses feature f_i.

By analyzing the feature importance scores, we can identify user-related features that play the most significant roles in predicting social media popularity. This finding underscores the importance of constructing a diverse set of hand-crafted features from the user's perspective to uncover latent indicators of post popularity. Notably, specific user-related features such as "Uid_post_diff_min", "Uid_post_diff_std", and "Uid_post_diff_mean", which capture metrics like minimum, standard deviation, and mean of user posts, emerge as particularly influential within this feature set.

4.6 Discussion

The research on SMP has made significant strides, particularly with the introduction of advanced multimodal representation learning techniques. However, several challenges and areas for future exploration remain. One primary challenge is the alignment of visual and textual feature spaces, which continues to be complex. While models like BLIP-2 demonstrate promise in bridging this gap, achieving seamless integration of visual and textual data remains crucial for enhancing the accuracy of popularity predictions. Effective alignment improves the semantic understanding of posts and addresses the limitations of traditional methods that often overlook the intricate interactions between different modalities.

In addition, the incorporation of user-related features is a critical aspect of SMP research. Our findings underscore the significant impact of user behavior and activity levels on post popularity. Historically, studies have focused primarily on content-based features; however, our research highlights the importance of dynamic user behavior patterns and activity levels. By analyzing sequences of tags used by users and their posting intervals, we can gain deeper insights into user interests and engagement levels. This approach not only improves prediction accuracy but also opens new avenues for personalized recommendations and targeted advertising. Future research could further explore integrating additional user-centric data, such as social network interactions and follower dynamics, to enhance predictive models further.

5 Conclusion

In this paper, we presented a novel Higher-order Vision-Language Alignment method for SMP, termed "HVLA". Our approach addresses the key challenges in predicting SMP by integrating multimodal data into an aligned semantic representation space and emphasizing user-related features. By leveraging the BLIP-2 model, we generated text-guided image descriptions that facilitate the seamless alignment of visual and textual information. Additionally, our detailed user behavior analysis provided deeper insights into post popularity dynamics. The experimental results demonstrated the effectiveness of our method, showing significant improvements in prediction accuracy. Future work can explore further enhancements by incorporating additional user interaction data and extending the model to handle other forms of social media content, such as videos and audio. Furthermore, investigating the temporal dynamics of post popularity in greater detail could provide more precise predictions and insights.

6 Acknowledgement

This work is supported by National Science and Technology Major Project (2023ZD0121101).

References

[1] Yongjie Cai, Hanghang Tong, Wei Fan, Ping Ji, and Qing He. 2015. Facets: Fast comprehensive mining of coevolving high-order time series. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 79–88.

[2] Weilong Chen, Chenghao Huang, Weimin Yuan, Xiaolu Chen, Wenhao Hu, Xinran Zhang, and Yanru Zhang. 2022. And-tag contrastive vision-and-language transformer for social media popularity prediction. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7008–7012.

[3] Weilong Chen, Chenghao Huang, Weimin Yuan, Xiaolu Chen, Wenhao Hu, Xinran Zhang, and Yanru Zhang. 2022. Title-and-Tag Contrastive Vision-and-Language Transformer for Social Media Popularity Prediction. In *Proceedings of the 30th ACM International Conference on Multimedia* (<conf-loc>, <city>Lisboa</city>, <country>Portugal</country>, </conf-loc>) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 7008–7012. https://doi.org/10.1145/3503161.3551568

[4] Xi Chen, Xiangmin Zhou, Jeffrey Chan, Lei Chen, Timos Sellis, and Yanchun Zhang. 2020. Event popularity prediction using influential hashtags from social media. *IEEE Transactions on Knowledge and Data Engineering* 34, 10 (2020), 4797–4811.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 4171–4186.

[6] Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Frontiers of Computer Science* 14 (2020), 241–258.

[7] Chih-Chung Hsu, Pi-Ju Tsai, Ting-Chun Yeh, and Xiu-Yu Hou. 2022. A comprehensive study of spatiotemporal feature learning for social medial popularity prediction. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7130–7134.

[8] Bogdan Ionescu, Maia Rohm, Bogdan Boteanu, Alexandru Lucian Gînscă, Mihai Lupu, and Henning Müller. 2020. Benchmarking image retrieval diversification techniques for social media. *IEEE Transactions on Multimedia* 23 (2020), 677–691.

[9] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. 2017. Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* 30 (2017).

[10] Cheng Li, Yue Lu, Qiaozhu Mei, Dong Wang, and Sandeep Pandey. 2015. Click-through prediction for advertising in twitter timeline. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1959–1968.

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven CH Hoi. 2022. BLIP: Bootstrapped Language-Image Pre-training for Unified Vision-Language Understanding and Generation. *arXiv preprint arXiv:2201.12086* (2022).

[12] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. 2017. UNIT: Unified Image and Text Representation Learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[13] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. VilBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *arXiv preprint arXiv:1908.02265* (2019).

[14] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems* 32 (2019).

[15] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*. 289–297.

[16] Shijian Mao, Wudong Xi, Lei Yu, Gaotian Lü, Xingxing Xing, Xingchen Zhou, and Wei Wan. 2023. Enhanced CatBoost with Stacking Features for Social Media Prediction. In *Proceedings of the 31st ACM International Conference on Multimedia* (Ottawa ON, Canada) *(MM '23)*. Association for Computing Machinery, New York, NY, USA, 9430–9435. https://doi.org/10.1145/3581783.3612839

[17] Mazin Mazloom, Martina Spagnuolo, Gianluigi Cornacchia, Vincent Oria, Roelof van Zwol, Paul Clough, and Benoit Huet. 2016. Multimodal popularity prediction of brand-related social media posts. In *Proceedings of the 2016 ACM on International Conference on Multimedia Retrieval*. 169–176.

[18] Swapnil Mishra. 2019. Bridging models for popularity prediction on social media. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 810–811.

[19] Elaheh Momeni, Claire Cardie, and Myle Ott. 2013. Properties, prediction, and prevalence of useful user-generated comments for descriptive annotation of social media objects. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7. 390–399.

[20] Author Name. 2024. Gemma: Advanced Language Model for Text Embedding. http://example.com/gemma Accessed: 2024-06-27.

[21] Yaroslav Nechaev, Francesco Corcoglioniti, and Claudio Giuliano. 2018. Type prediction combining linked open data and social media. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*. 1033–1042.

[22] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[23] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, and Andrey Gulin. 2018. CatBoost: unbiased boosting with categorical features. *Advances in neural information processing systems* 31 (2018).

[24] Xueming Qian, He Feng, Guoshuai Zhao, and Tao Mei. 2013. Personalized recommendation combining user interest and social circle. *IEEE transactions on knowledge and data engineering* 26, 7 (2013), 1763–1777.

[25] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pam Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*.

[26] Miriam Redi and Georges Quenot. 2022. Overview of the Multimedia Grand Challenges 2022. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7220–7222.

[27] Ramit Sawhney, Shivam Agarwal, Arnav Wadhwa, and Rajiv Shah. 2020. Deep attentive learning for stock movement prediction from social media text and company correlations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8415–8426.

[28] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).

[29] Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5100–5111.

[30] Jiliang Tang, Shiyu Chang, Charu Aggarwal, and Huan Liu. 2015. Negative link prediction in social media. In *Proceedings of the eighth ACM international conference on web search and data mining*. 87–96.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[32] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[33] Dashun Wang, Mohammad Mazloom, Madian Khabsa, Max Berzofsky, and Jie Tang. 2015. Learning to discover social circles in ego networks. In *Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 399–404.

[34] Kai Wang, Penghui Wang, Xin Chen, Qiushi Huang, Zhendong Mao, and Yongdong Zhang. 2020. A feature generalization framework for social media popularity prediction. In *Proceedings of the 28th ACM international conference on multimedia*. 4570–4574.

[35] Bo Wu, Wen-Huang Cheng, Peiye Liu, Bei Liu, Zhaoyang Zeng, and Jiebo Luo. 2019. Smp challenge: An overview of social media prediction challenge 2019. In *Proceedings of the 27th ACM International Conference on Multimedia*. 2667–2671.

[36] Bo Wu, Wen-Huang Cheng, Yongdong Zhang, Huang Qiushi, Li Jintao, and Tao Mei. 2017. Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks. In *International Joint Conference on Artificial Intelligence (IJCAI)*.

[37] Bo Wu, Peiye Liu, Wen-Huang Cheng, Bei Liu, Zhaoyang Zeng, Jia Wang, Qiushi Huang, and Jiebo Luo. 2023. SMP Challenge: An Overview and Analysis of Social Media Prediction Challenge. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9651–9655.

[38] Bo Wu, Peiye Liu, Wen-Huang Cheng, Bei Liu, Zhaoyang Zeng, Jia Wang, Qiushi Huang, and Jiebo Luo. 2023. SMP Challenge: An Overview and Analysis of Social Media Prediction Challenge. In *Proceedings of the 31th ACM International Conference on Multimedia*.

[39] Jianmin Wu, Liming Zhao, Dangwei Li, Chen-Wei Xie, Siyang Sun, and Yun Zheng. 2022. Deeply Exploit Visual and Language Information for Social Media Popularity Prediction. In *Proceedings of the 30th ACM International Conference on Multimedia* (Lisboa, Portugal) *(MM '22)*. Association for Computing Machinery, New York, NY, USA, 7045–7049. https://doi.org/10.1145/3503161.3551576

[40] Jianmin Wu, Liming Zhao, Dangwei Li, Chen-Wei Xie, Siyang Sun, and Yun Zheng. 2022. Deeply exploit visual and language information for social media popularity prediction. In *Proceedings of the 30th ACM International Conference on Multimedia*. 7045–7049.

[41] Kele Xu, Zhimin Lin, Jianqiao Zhao, Peicang Shi, Wei Deng, and Huaimin Wang. 2020. Multimodal deep learning for social media popularity prediction with attention mechanism. In *Proceedings of the 28th ACM International Conference on Multimedia*. 4580–4584.

[42] Yongquan Yang, Haijun Lv, and Ning Chen. 2023. A survey on ensemble learning under the era of deep learning. *Artificial Intelligence Review* 56, 6 (2023), 5545–5589.